Multicenter Validation of Individual Preoperative Motor Outcome Prediction for Deep Brain Stimulation in Parkinson's Disease

Jeroen G.V. Habets^a Christian Herff^a Alfonso A. Fasano^{b, c} Martijn Beudel^d
Ersoy Kocabicak^e Alfons Schnitzler^f Muneer Abu Snineh^g Suneil K. Kalia^{b, c}
Carolina Ramirez-Gómez^{b, c} Mojgan Hodaie^{b, c, h} Renato P. Munhoz^{b, c}
Eline Rouleau^d Onur Yildiz^e Eduard Linetsky^g Rick Schuurmanⁱ
Christian J. Hartmann^f Andres M. Lozano^{b, c} Rob M.A. De Bie^d Yasin Temel^a
Marcus L.F. Janssen^j

^aDepartment of Neurosurgery, School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, The Netherlands; ^bEdmond J. Safra Program in Parkinson's Disease, Morton and Gloria Shulman Movement Disorders Clinic, Toronto Western Hospital, UHN, Krembil Brain Institute, Toronto, ON, Canada; ^cDivision of Neurology, University of Toronto, Toronto, ON, Canada; ^dDepartment of Neurology, Amsterdam Neuroscience Institute, Amsterdam University Medical Center, Amsterdam, The Netherlands; ^eNeuromodulation Center and Department of Neurosurgery, Ondokuz Mayıs University, Samsun, Turkey; ^fDepartment of Neurology, Institute of Clinical Neuroscience and Medical Psychology, Centre for Movement Disorders and Neuromodulation, Medical Faculty, Universitätsklinikum Duesseldorf, Duesseldorf, Germany; ^gDepartment of Neurology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel; ^hDivision of Neurosurgery, University Health Network and Department of Surgery, University of Toronto, Toronto, ON, Canada; ⁱDepartment of Neurosurgery, Amsterdam UMC, Amsterdam, The Netherlands; ^jDepartment of Neurology and Clinical Neurophysiology, School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, The Netherlands

Keywords

Parkinson's disease · Deep brain stimulation · Motor response · Outcome prediction · Machine learning · Clinical prediction models

Abstract

Background: Subthalamic nucleus deep brain stimulation (STN DBS) is an established therapy for Parkinson's disease (PD) patients suffering from motor response fluctuations despite optimal medical treatment, or severe dopaminergic side effects. Despite careful clinical selection and surgical procedures, some patients do not benefit from STN DBS. Pre-

operative prediction models are suggested to better predict individual motor response after STN DBS. We validate a pre-registered model, DBS-PREDICT, in an external multicenter validation cohort. **Methods:** DBS-PREDICT considered eleven, solely preoperative, clinical characteristics and applied a logistic regression to differentiate between weak and strong motor responders. Weak motor response was defined as no clinically relevant improvement on the Unified Parkinson's Disease Rating Scale (UPDRS) II, III, or IV, 1 year after surgery, defined as, respectively, 3, 5, and 3 points or more. Lower UPDRS III and IV scores and higher age at disease onset contributed most to weak response predictions. Individual predictions were compared with actual clinical outcomes. **Re-**

ults: 322 PD patients treated with STN DBS from 6 different centers were included. DBS-PREDICT differentiated between weak and strong motor responders with an area under the receiver operator curve of 0.76 and an accuracy up to 77%.

Conclusion: Proving generalizability and feasibility of preoperative STN DBS outcome prediction in an external multi-center cohort is an important step in creating clinical impact in DBS with data-driven tools. Future prospective studies are required to overcome several inherent practical and statistical limitations of including clinical decision support systems in DBS care.

© 2021 The Author(s)

Published by S. Karger AG, Basel

Introduction

Parkinson's disease (PD) is a neurodegenerative disease characterized by motor and non-motor symptoms and will affect up to 12 million patients worldwide by 2040 [1, 2]. While oral dopamine replacement therapies treat early disease symptoms well, disease progression or wearing-off effects may cause motor and non-motor fluctuations over time [3]. Subthalamic nucleus (STN) deep brain stimulation (DBS) is an accepted and proven symptomatic treatment of early and advanced motor response fluctuations with satisfactory overall improvement in the majority of patients [4–6]. However, up to one-third experiences suboptimal motor improvement postoperatively [7, 8], and decreasing this amount could lead to clinical and socioeconomic impact. Several strategies are suggested to improve motor response such as improving preoperative selection [9, 10], optimizing surgical lead

placement [11], and optimizing postoperative DBS programming or stimulation paradigms [12–14].

To create clinical and socioeconomic impact via preoperative selection, the relation between preoperative variables and postoperative outcome must be strongly predictive despite the absence of surgical factors, such as lead positioning, which are unknown in the preoperative phase. Preoperative significant quality of life (QoL) impairment, severe motor symptoms, greater levodopa response, better cognitive performance, tremor-dominancy, and younger age have been correlated with better postoperative motor function improvement and QoL [15–19]. It is important to note that there is no consensus about all relevant preoperative predictors, and some of them specifically are a matter of debate [20]. Unlike the studies reporting these correlations [16–19], 2 studies developed a machine learning model to decrease the number of suboptimal responders with individual motor response predictions (Fig. 1) [21, 22]. Both proof-of-concept studies described a good prediction of a binary outcome in a retrospective single-center STN DBS cohort. They both were designed for preoperative application and did not consider intra- or postoperative variables.

Frizon et al. [21] defined favorable outcome as clinically relevant QoL improvement, and their model considered 3 preoperative variables which were most associated with favorable outcome. A logistic regression model was applied and validated using bootstrap sampling. The model applied in this study, DBS-PREDICT, defined unfavorable outcome as no clinically relevant improvement on the Unified Parkinson's Disease Rating Scale (UPDRS) II, III, and IV 1 year after DBS implantation (Fig. 2). This

Fig. 1. Schematic representation of methodology. Individual preoperative variables from 6 international centers are used as validation cohort (left panel). DBS-PREDICT's preregistered logistic regression model is trained on a separate development cohort which is described earlier. It takes preoperative variables from one individual patient as input (middle panel). DBS-PREDICT returns an individual probability between 0 and 1 for becoming a weak motor responder 1 year after STN DBS (red arrow, right panel). The probability results in a "weak response" prediction if the probability is larger than the predefined threshold (orange dotted line). STN DBS, subthalamic nucleus deep brain stimulation.

outcome classification is strict as it only requires improvement on one of the 3 variables to claim favorable outcome. DBS-PREDICT also utilized a logistic regression model which considered all 15 available preoperative variables and was validated via a 10-fold cross-validation [22]. To enable model generalizability and statistical transparency, DBS-PREDICT was retrained without preoperative neuropsychological and UPDRS I scores and preregistered before data collection completion (ClinicalTrials.gov, NCT04093908) [23]. Performance and predicting variable importance did not change relevantly after retraining (Fig. 3; online suppl. material; for all online suppl. material, see www.karger.com/doi/10.1159/000519960).

To develop valid and generalizable clinical prediction models with an impactful clinical utilization, external validation is needed after model development and proof of concept [24, 25]. Here, we present an external multicenter validation of the preregistered prediction model DBS-PREDICT. Additionally, remaining challenges for impactful DBS outcome prediction tools are discussed, such as the design of follow-up validation studies, the requirements and limitations of outcome definition, and the intended clinician-prediction tool interaction.

Methods

DBS-PREDICT Development

Development details of the model are described previously [22] and will be briefly summarized here. In the development study, all available preoperative predictors were included without preselecting predictors based on a sensitivity analysis. The preoperative variables included are gender, age at PD onset, PD duration, age at DBS, levodopa equivalent daily dosage (LEDD), UPDRS I, II, III, and IV, Hoehn and Yahr scales, and relevant neuropsychological assessments (Stroop test interference and verbal fluency categorical and letter tests). All preoperative tests were reported in on-medication condition; only UPDRS III and Hoehn and Yahr scale were reported both in on- and off-medication condition.

Strong motor response was defined as a minimal clinical important difference (MCID) on UPDRS II, or III, or IV 1 year after STN DBS implementation (Fig. 2). Based on literature, MCIDs were set as, respectively, 3, 5, and 3 points [26–30]. We chose to consider only UPDRS III change in on-stimulation and on-medication condition, compared to preoperative on-medication because the daily life reflection of the fully dopamine-deprived off-medication condition can be disputed. Due to the daily life influence of off-medication moments, we aim to capture via the narrative items regarding motor aspects of experiences of daily living in UPDRS II and regarding off-fluctuations in UPDRS IV. This definition results in weak responders who do not show an MCID in activities of daily life (UPDRS II), nor in motor symptoms in optimal therapy condition (UPDRS III), and nor in the burden or duration of adverse effects and off-periods (UPDRS IV).

DBS-PREDICT generates for every individual a weak responder probability, which is translated into the binary outcome based on a threshold of probability acceptance (Fig. 1). The proof-of-concept analysis demonstrated an area under the curve (AUC) of the receiver operator curve of 0.88 (± 0.14) and a classification accuracy of 78% [22]. During the initial model development, which was used for preregistration, optimal predictive performance was achieved by accepting a threshold above 0.24 for weak responders [23, 31].

The weighing of preoperative variables driving the prediction and the classification accuracy did not show relevant changes after the exclusion of the neuropsychological variables and UPDRS I scores for generalizability (see description in online suppl. material II, Fig. 3 and online suppl. Fig. 2, 3). The model is made available on GitHub (<https://github.com/jgvhabets/DBSPREDICT>).

Fig. 2. Flowchart demonstrating the definition of strong and weak motor response. Only if no clinically relevant improvement in UPDRS II, III, or IV is found, a patient is considered to be a weak responder. UPDRS: Unified Parkinson Disease Rating Scale.

Fig. 3. Importance of all available preoperative predictive variables in DBS-PREDICT's logistic regression analysis. Importances are expressed in normalized relative weights. Relative weights are retrieved from DBS-PREDICT's logistic regression function after training the model on the development cohort. UPDRS: Unified Parkinson Disease Rating Scale; HY, Hoehn and Yahr scale; DBS, deep brain stimulation; LEDD, levodopa equivalent daily dosage.

Patient Population

A retrospective multicenter cohort was collected as validation dataset. DBS centers were invited to provide data from PD patients who were treated with STN DBS. The following preoperative variables were collected: age at PD onset, age at DBS, disease duration, gender, LEDD, preoperative UPDRS II, III in on-medication and off-medication conditions, and IV scores, and HY scales in on- and off-medication conditions. The following postoperative variables were collected from 1-year follow-up moments: UPDRS II, III in on-medication and on-stimulation conditions, and IV scores. Approval of an Ethical Committee was received that this study agreed with the Good Clinical Practice norms and the Declaration of Helsinki (Medical Ethical Testing Committee azM/UM, Maastricht, The Netherlands, record: 2018-0739-A-9).

Descriptive Statistical Methods

Pre- and postoperative variables from the current multicenter population and the initial single-center training population were compared. Continuous variables were evaluated for significant differences with Mann-Whitney U analyses since sample sizes differed and not all variables were normally distributed. The distributions of weak responders in the multicenter validation cohort and the development cohort were compared using a χ^2 test.

Predictive Statistical Methods

DBS-PREDICT translated the preoperative variables into individual probabilities to become a weak responder. The generated individual probabilities were weighted against predefined thresholds, resulting in individual binary outcome predictions (Fig. 1). The predictive performance was analyzed with both a default probability acceptance threshold of 0.5 [32] as well as the optimal threshold from the development study, 0.24, since the development cohort was as unbalanced as the validation cohort [22]. To provide a valid and understandable evaluation of the performance of the model, we report the AUC of the receiver operator curve, positive and negative predictive values (PPV and NPV, respectively), the sensitivity, and the accuracy. Findings are presented in line with TRIPOD guidelines for artificial intelligence predictive modeling [33, 34].

Results

Patient Population

Six different DBS centers in Europe, Asia, and North America participated in this validation study. In total, 344 PD patients were included who were treated with STN DBS, completed a 1-year postoperative follow-up, and whose records contained all required variables. After removing missing data, 322 patients were included in the validation cohort. The descriptive statistics and distributions of the pre- and postoperative variables of the validation cohort are shown and compared with the development cohort in online supplementary Table 1.

The presented validation cohort had 26% weak responders (83 out of 322), compared to 30% in the development cohort (27 out of 90). This indicated that the populations do not have statistically significant differences in distribution of strong and weak responders (χ^2 statistic of 2.892, $p = 0.089$). To test statistical significance of individual pre- and postoperative variables, a Bonferroni-corrected p value of 0.0038 is used. The validation cohort had a lower rate of female patients. The preoperative UPDRS III scores were significantly higher in the development cohort, together with a significantly lower LEDD preoperatively. Postoperatively, both UPDRS III and LEDD did not differ significantly between the 2 cohorts.

Predictive Performance

The individual preoperative variables of all 322 patients were inserted in DBS-PREDICT and this resulted in 322 individual weak responder probabilities between 0

Fig. 4. Performance of prediction model in validation cohort. Left: ROC showing true positive rates and false positive rates for different thresholds; both analyzed thresholds are showed with red circles. Middle: confusion matrix corresponding to 0.50 threshold. Right: confusion matrix corresponding to 0.24 threshold. ROC, receiver operator curve.

and 1. All individual probabilities were weighted against the 2 different preregistered thresholds, leading to 322 binary outcome predictions per threshold which were compared with the binary outcome classes based on the actual postoperative variables.

The first threshold was a default probability acceptance of 0.50 [32]. The second threshold was optimal from the development study, 0.24 [23, 31].

DBS-PREDICT differentiated between strong and weak motor responders with an AUC of 0.76 (Fig. 4, left panel). The default threshold, accepting every probability above 0.50 [32], led to an accuracy of 0.77 (248 correct predictions out of 322, Fig. 4 middle panel). The numbers in the confusion matrix led to a sensitivity of 0.29 (predicted 24 out of 83 true weak responders correctly), a specificity of 0.94 (predicted 224 out of 239 true strong responders correctly), a PPV of 0.62 (out of 39 weak predictions, 24 were correct), and an NPV of 0.79 (out of 283 strong predictions, 224 were correct) (Table 1). The threshold based on the development cohort's optimum 0.24 [23, 31] led to an accuracy of 0.73 (236 correct predictions out of 322, Fig. 4, right panel). This threshold led to a sensitivity of 0.58 (predicted 48 out of 83 true weak responders correctly), a specificity of 0.79 (predicted 188 out of 239 true strong responders correctly), a PPV of 0.49 (out of 99 weak predictions, 48 were correct), and an NPV of 0.84 (out of 223 strong predictions, 188 were correct) (Table 1).

Table 1. Predictive performance of DBS-PREDICT for 2 analyzed thresholds for accepting probabilities to become a weak responder

	Threshold 0.5	Threshold 0.24 (from development results)
Accuracy	0.77 (248/322)	0.73 (236/322)
Sensitivity	0.29 (24/83)	0.58 (48/83)
Specificity	0.94 (224/239)	0.79 (183/239)
PPV	0.62 (24/39)	0.49 (48/99)
NPV	0.79 (224/283)	0.84 (188/223)

PPV, positive predictive value; NPV, negative predictive value.

Discussion

We demonstrated the feasibility and generalizability of the preregistered DBS-PREDICT model in a retrospective multicenter validation cohort which is representative for the global STN DBS population. Our results are an important next step toward impactful clinical decision support systems (CDSSs) for DBS care. They underline the potential of preoperative individual motor response prediction as a method to improve STN DBS motor outcome [21, 22]. In general, the translation from a potent machine learning prediction model into a CDSS with clinical impact is notoriously complex and faces several inherent challenges [35, 36]. Therefore, our findings and considerations are valuable for DBS

CDSS development regardless of input variables or DBS indication.

Clinical impact of a CDSS is defined as the change in patient care resulting from its implementation [24] and depends on the following 3 factors: (1) the generalizability of the model to external patient populations and centers [24, 25], (2) the quality and understandability of numeric predictive results of a model [24], and (3) the presence of a well-considered, clinically realistic utilization, including the socioeconomic impact of the utilization [37]. We will discuss how our results contribute to points 1 and 2, which limitations should be considered, and how this work paves the way for investigating point 3 [25].

Model Generalizability and Usability

Generalizability of the preregistered DBS-PREDICT model was demonstrated by the high accuracy and AUC score in the global external validation cohort. The descriptive variables of this cohort including the proportion of weak responders (26%) were comparable with and representative for global STN DBS populations (online suppl. Table 1) [4–7]. The development and validation cohorts showed high similarity regarding descriptive variables, and the described differences in preoperative characteristics were interpreted as not troublesome (online suppl. Table 1). The most relevant difference between the 2 cohorts was the higher preoperative UPDRS III scores in the development cohort together with the lower amount of LEDD (online suppl. Table 1). This could be explained by a more conservative pharmacological strategy in the development cohort compared to the validation cohort.

To ensure generalizability and usability among different DBS centers, DBS-PREDICT is deliberately founded on preoperative, clinical variables which do not require specific radiologic or neurophysiological features. It is beyond doubt that perioperative and postoperative, surgical, radiological, and neurophysiological variables influence a patient's motor response. DBS-PREDICT aims nonetheless to provide a numeric support about expected motor response during preoperative counseling, within the possibilities of this conceptual limitation. Also, more detailed clinical variables such as UPDRS subscores instead of sum scores have the theoretical potential to improve individualization of the prediction model. In a retrospective setting, it is difficult to include large enough populations including similar detailed scores. Therefore, we suggest preregistered, prospective data collection to investigate the additional value of detailed clinical variables.

Quality, Understandability, and Interpretation of Predictive Performance

DBS-PREDICT showed a good classification accuracy in both analyzed utilizations of 0.77 and 0.73 and an AUC of 0.76 (Fig. 4; Table 1). The absence of a comparable validation study of a DBS outcome CDSS in the literature does not allow comparison of these results. However, results are in line with the 2 available model development studies [21, 22]. Overall, this supports the general feasibility of preoperative identification of weak motor responders in STN DBS care. Nevertheless, for an exact clinical utilization of DBS-PREDICT, the performance must be interpreted in more detail. Despite the high specificity scores, the current positive predictive values and sensitivity scores are not sufficient to withdraw predicted weak responders from DBS implantation (Table 1). This underlines that a CDSS for a highly complex and multifactorial process as DBS candidate selection should rather play a more cautious supportive and advisory role, than replacing the team of clinicians in decision-making [24, 38]. The balance between sensitivity and specificity and PPV and NPV together with the clinical definition of “positive” and “negative” should influence the role of the CDSS in clinical practice.

The applied logistic regression, intentionally chosen over more complex machine learning models, contributes to this interpretability among clinicians. In general, logistic regression models are often non-inferior compared to more complex models for clinical applications [39].

To improve both the predictive performance and the clinical understanding among clinicians, future CDSS development could combine clinical judgment and numerical support by a CDSS. For example, such a “hybrid-CDSS” could be aimed at a population selected by the clinician, or the numerical strengths and weaknesses of a CDSS could decide where human-clinician input is necessary. This would also allow us to explore the interaction between clinician and CDSS.

Optimal Target Patients and Timing of an Impactful CDSS

The current development and validation cohorts of DBS-PREDICT included patients included for STN DBS based on the current clinical selection. In future studies, all DBS candidates could be included, for example, in a validation study to test whether a CDSS confirms the clinically rejected candidates as weak responders. DBS candidates eventually rejected based on surgical exclusion criteria not covered by the CDSS variables should not be included in these analyses.

The exact moment of use in clinical DBS practice will also influence the interpretation of a CDSS's balance between false positive and false negative prediction. The consequences of a weak responder falsely predicted to respond strong, and vice versa, are depending on the CDSS's role in clinical decision-making. An open and understandable explanation from CDSS-developers toward clinicians and patients on implications of sensitivity, specificity, and PPV versus NPV is of utmost importance to create confidence in machine learning outcome predictions and to realize clinical impact through CDSS [24, 36, 38]

The Potential Socioeconomic Impact

DBS care in PD is subject to large variations in disease symptomatology and progression, therapeutic strategies, and international differences in insurance and health care systems. These factors complicate a valid indication or measurement of the potential socioeconomic impact of a DBS outcome CDSS. Current DBS therapy in PD is cost-effective compared to best medical treatment. This means the increased QoL is relatively larger than the increased health care costs [40]. Improved preoperative counseling and selection with CDSSs hold potential to lower absolute perioperative costs by preventing unsuccessful cases. Moreover, patients experiencing a less favorable outcome might be better prepared for this due to the improved preoperative counseling. This could lead to increased postoperative satisfaction, resulting in a better QoL [41]. Thorough socioeconomic analyses should be included in future prospective CDSS development and data collection [25]. When a CDSS is not proven yet, data collection without interfering in clinical care should be considered [25].

Alternative Approaches for Preoperative STN DBS Outcome Improvement

We want to stress that preoperative prediction based on clinical variables is one of several approaches to improve patient outcome. Advances in structural and functional imaging of the whole brain and its connectivity hold potential to improve the preoperative target selection and surgical planning [9]. Better understanding of the STN itself can lead to more stable motor improvements as well [11]. Moreover, increasing insights in the clinical non-motor characteristics, genetic profiles, and neurophysiological characteristics and their relation with DBS outcome have promising potential to improve patient selection, counseling, and outcome prediction [18, 42]. All these approaches do not exclude each other

and can improve DBS outcome in PD in a complementary way in the future. The discussed considerations regarding clinical impact realization, outcome definition, and future study design are relevant for all different DBS CDSSs.

Limitations

This study is subject to several limitations, of which some are inherent to its retrospective nature. Limitations are minimized by preregistering the model, which mimicked a prospective application of the model. While the preregistration strengthened the statistical value, it did not allow the inclusion of new variables, such as QoL scores and detailed non-motor symptomatology. We attempted to overcome the lack of a QoL instrument as an outcome variable, by using a holistic outcome classification. The outcome classification covers MCIDs in a wide spectrum of factors influencing QoL: daily life performance, motor symptom severity, and adverse effects suffering (UPDRS II, III, and IV, respectively). To ensure generalizability, only sum scores were included instead of detailed clinical variables. Variance in clinical and administrative processes only allows a complete and detailed inclusion of pre- and postoperative variables via prospective data collection. Obviously, a more complete and detailed data collection would require a development study first and an external validation study afterward.

The binary outcome classification can be disputed. This approach, however, aligns with the clinical decision in the individual patient to offer DBS or not. We emphasize that we followed large DBS trials in the pragmatic choice of selecting UPDRS sum scores as outcome variables [7, 8]. For the evidence and considerations we provide here, consensus about the outcome definition does not have the highest priority. For future prospective, pre-registered, DBS CDSS data collection, this has one of the highest priorities.

Local variation in pharmacological and DBS-programming strategies are neglected by the model. Due to the amount of data required by data-driven models, this currently cannot be overcome, and a wide generalizability is essential. Further, preoperative prediction of DBS outcome will always be limited by the absence of peri- and postoperative variables describing surgical electrode placement and surgical complications. This limitation is inherent to clinical role of a preoperative CDSS. Postoperative approaches to improve DBS outcome like directional steered stimulation or closed-loop DBS should always be considered for a suboptimal responding patient group.

Conclusion

Preoperative individual STN DBS outcome prediction is feasible in an external multicenter cohort using a pre-registered model. The results and perspectives endorse next steps in the development of impactful CDSSs for individualized DBS care.

We emphasize the need of prospective and preregistered multicenter studies to overcome clinical, practical, and statistical challenges in the development of CDSS models for individual DBS decision-making in PD care. These studies can be considered an obligatory next step before clinical implementation.

Statement of Ethics

The study was approved in agreement with the Good Clinical Practice norms and the World Medical Declaration of Helsinki by the Medical Ethical Testing Committee azM/UM (Maastricht, The Netherlands) under record 2018-0739-A-9, including an exemption of individual informed consent due to the anonymity and the lack of traceability of the collected data.

Conflict of Interest Statement

None of the authors has a conflict of interest.

Funding Sources

No direct funding is received and there is no conflict of interest due to the reported funding for this work. RMAB received research support from Netherlands Organization for Health Research and Development, Parkinson Vereniging, Stichting Parkinson Nederland, GE Health, Medtronic, Lysosomal Therapeutics, Neuroderm, all paid to the department. AAF received research support from Medtronic and Boston Scientific and honoraria from Abbvie, Abbott, Cerecare, Boston Scientific, and Medtronic. CJH received honoraria from Abbott. MH received honoraria from Medtronic. SKK is consultant for Medtronic for declaration. AML is consultant for Medtronic, Boston Scientific, and Abbott. MAS received consultation fees from Abbvie and research support from Medtronic and Boston Scientific. RS is consultant for Medtronic, Boston Scientific, and Elekta. YT received a research grant from Weijerhorst Foundation. CH received funding from Dutch Research Council (NWO).

Author Contributions

J.H., C.H., M.J., and Y.T. contributed to research conception. J.H., C.H., M.J., Y.T., M.B., A.F., A.S., M.A.S., and E.K. contributed to the organization. J.H. and C.H. contributed to the statistical design: J.H. drafted the manuscript. All authors contributed to the execution, review, and critique.

Data Availability Statement

All anonymized data, except from one center due to local ethical and legal restrictions, will be made available upon reasonable request. The prediction model is made available on <https://github.com/jgvhabets/DBSPREDICT>.

References

- 1 Titova N, Padmakumar C, Lewis SJG, Chaudhuri KR. Parkinson's: a syndrome rather than a disease? *J Neural Transm*. 2017 Aug;124(8):907–14.
- 2 Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. *J Parkinsons Dis*. 2018;8(s1):S3–8.
- 3 Kalia LV, Lang AE. Parkinson's disease. *Lancet*. 2015 Aug 29;386(9996):896–912.
- 4 Odekerken VJ, van Laar T, Staal MJ, Mosch A, Hoffmann CF, Nijssen PC, et al. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. *Lancet Neurol*. 2013 Jan;12(1):37–44.
- 5 Schuepbach WM, Rau J, Knudsen K, Volkmann J, Krack P, Timmermann L, et al. Neurostimulation for Parkinson's disease with early motor complications. *N Engl J Med*. 2013 Feb 14;368(7):610–22.
- 6 Janssen ML, Duits AA, Turaihi AH, Ackermans L, Leentjens AF, Leentjens AF, et al. Subthalamic nucleus high-frequency stimulation for advanced Parkinson's disease: motor and neuropsychological outcome after 10 years. *Stereotact Funct Neurosurg*. 2014;92(6):381–7.
- 7 Deuschl G, Schade-Brittinger C, Krack P, Volkmann J, Schäfer H, Bötzel K, et al. A randomized trial of deep-brain stimulation for Parkinson's disease. *N Engl J Med*. 2006 Aug 31;355(9):896–908.
- 8 Williams A, Gill S, Varma T, Jenkinson C, Quinn N, Mitchell R, et al. Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. *Lancet Neurol*. 2010 Jun;9(6):581–91.
- 9 Bari AA, Fasano A, Munhoz RP, Lozano AM. Improving outcomes of subthalamic nucleus deep brain stimulation in Parkinson's disease. *Expert Rev Neurother*. 2015 Oct 03;15(10):1151–60.
- 10 Hartmann CJ, Fliegen S, Groiss SJ, Wojtecki L, Schnitzler A. An update on best practice of deep brain stimulation in Parkinson's disease. *Ther Adv Neurol Disord*. 2019;12:1756286419838096.
- 11 Bot M, Schuurman PR, Odekerken VJJ, Verhagen R, Contarino FM, De Bie RMA, et al. Deep brain stimulation for Parkinson's disease: defining the optimal location within the subthalamic nucleus. *J Neurol Neurosurg Psychiatry*. 2018 May;89(5):493–8.
- 12 Picillo M, Lozano AM, Kou N, Puppi Munhoz R, Fasano A. Programming deep brain stimulation for Parkinson's disease: the Toronto western hospital algorithms. *Brain Stimul*. 2016 May-Jun;9(3):425–37.
- 13 Pina-Fuentes D, Little S, Oterdoom M, Neal S, Poghosyan A, Tijssen MAJ, et al. Adaptive DBS in a Parkinson's patient with chronically implanted DBS: a proof of principle. *Mov Disord*. 2017 Aug;32(8):1253–4.

- 14 Schnitzler ABM, Verhagen L, Groppa S, Cheeran B, Karst E, Defresne F, et al. [Directional versus omnidirectional deep brain stimulation for Parkinson's disease: results of a prospective, blinded, multi-center, single-arm crossover study \[abstract\]](#). Nice, France: Movement Disorders Society Conference; 2019.
- 15 Cavallieri F, Fraix V, Bove F, Mulas D, Tondelli M, Castrioto A, et al. Predictors of long-term outcome of subthalamic stimulation in Parkinson disease. *Ann Neurol*. 2021 Mar; 89(3):587–97.
- 16 Daniels C, Krack P, Volkman J, Raethjen J, Pinsker MO, Kloss M, et al. Is improvement in the quality of life after subthalamic nucleus stimulation in Parkinson's disease predictable? *Mov Disord*. 2011 Dec;26(14):2516–21.
- 17 Abboud H, Genc G, Thompson NR, Oravittanakul S, Alsallom F, Reyes D, et al. Predictors of functional and quality of life outcomes following deep brain stimulation surgery in Parkinson's Disease patients: disease, patient, and surgical factors. *Parkinsons Dis*. 2017;2017:5609163.
- 18 Dafsari HS, Weiss L, Silverdale M, Rizos A, Reddy P, Ashkan K, et al. Short-term quality of life after subthalamic stimulation depends on non-motor symptoms in Parkinson's disease. *Brain Stimul*. 2018 Jul-Aug;11(4):867–74.
- 19 Schuepbach WMM, Tonder L, Schnitzler A, Krack P, Rau J, Hartmann A, et al. Quality of life predicts outcome of deep brain stimulation in early Parkinson disease. *Neurology*. 2019 Feb 8;92(24):2.
- 20 Geraedts VJ, Feleus S, Marinus J, van Hilten JJ, Contarino MF. What predicts quality of life after subthalamic deep brain stimulation in Parkinson's disease? A systematic review. *Eur J Neurol*. 2020 Mar;27(3):419–28.
- 21 Frizon LA, Hogue O, Achey R, Floden DP, Nagel S, Machado AG, et al. Quality of life improvement following deep brain Parkinson's disease: development of a prognostic model. *Neurosurgery*. 2019 Sep 1;85(3):343–9.
- 22 Habets JGV, Janssen MLF, Duits AA, Sijben LCJ, Mulders AEP, De Greef B, et al. Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease: proof of principle in a retrospective cohort. *PeerJ*. 2020 Nov 18;8:e10317.
- 23 ClinicalTrials.gov. NCT04093908: Prediction of STN DBS Motor Response in PD (DBS-PREDICT). 2019. Available from: clinicaltrials.gov.
- 24 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019 Oct 29;17(1):195.
- 25 Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models - development, evaluation, and clinical application. *N Engl J Med*. 2020 Apr 23;382(17):1583–6.
- 26 Schrag A, Sampaio C, Counsell N, Poewe W. Minimal clinically important change on the unified Parkinson's disease rating scale. *Mov Disord*. 2006 Aug;21(8):1200–7.
- 27 Shulman LM, Gruber-Baldini AL, Anderson KE, Fishman PS, Reich SG, Weiner WJ. The clinically important difference on the unified Parkinson's disease rating scale. *Arch Neurol*. 2010 Jan;67(1):64–70.
- 28 Horvath K, Aschermann Z, Acs P, Deli G, Janszky J, Komoly S, et al. Minimal clinically important difference on the motor examination part of MDS-UPDRS. *Parkinsonism Relat Disord*. 2015 Dec;21(12):1421–6.
- 29 Kostoglou K, Michmizos KP, Stathis P, Sakas D, Nikita KS, Mitsis GD. Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings. *IEEE Trans Biomed Eng*. 2017 May;64(5):1123–30.
- 30 Makkos A, Kovács M, Aschermann Z, Harmat M, Janszky J, Karádi K, et al. Are the MDS-UPDRS-based composite scores clinically applicable? *Mov Disord*. 2018 May; 33(5):835–9.
- 31 Habets Jd A, Sijben L, De Greef B, Mulders A, Temel Y, Kuijff M, et al. [pre-print] Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease. *MedRxiv*. 2019.
- 32 Zou Q, Xie S, Lin Z, Wu M, Ju Y. Finding the best classification threshold in imbalanced classification. *Big Data Res*. 2016 Sep 01;5:2–8.
- 33 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;122(1):434.
- 34 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019 Apr 20;393(10181):1577–9.
- 35 Mateen BA, Liley J, Denniston AK, Holmes CC, Vollmer SJ. Improving the quality of machine learning in health applications and clinical research. *Nat Mach Intell*. 2020 Oct 01; 2(10):554–6.
- 36 Pedersen M, Verspoor K, Jenkinson M, Law M, Abbott DF, Jackson GD. Artificial intelligence for clinical decision support in neurology. *Brain Commun*. 2020;2(2):fcaa096.
- 37 Shah NH, Milstein A, Bagley PDS. Making machine learning models clinically useful. *Jama*. 2019 Oct 8;322(14):1351–2.
- 38 Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med*. 2018 Aug 28;1(1):40.
- 39 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019 Jun;110:12–22.
- 40 Pietzsch JB, Garner AM, Marks WJ Jr. Cost-effectiveness of deep brain stimulation for advanced Parkinson's disease in the United States. *Neuromodulation*. 2016 Oct;19(7): 689–97.
- 41 Lin HY, Hasegawa H, Mundil N, Samuel M, Ashkan K. Patients expectations and satisfaction in subthalamic nucleus deep brain stimulation for Parkinson disease: 6-year follow-up. *World Neurosurg*. 2019 Jan;121:e654–e60.
- 42 Jost ST, Visser-Vandewalle V, Rizos A, Loehrer PA, Silverdale M, Evans J, et al. Non-motor predictors of 36-month quality of life after subthalamic stimulation in Parkinson disease. *NPJ Parkinsons Dis*. 2021 Jun 08;7(1): 48.